Verso una (open) peer review dei dati: uno studio pilota nelle scienze sociali

Daniela Luzi, Roberta Ruggieri, Lucio Pisacane, Rosa Di Cesare
Consiglio Nazionale delle Ricerche. Istituto di Ricerche sulla Popolazione e le Politiche Sociali, Roma

1. INTRODUZIONE

La condivisione dei dati della ricerca viene enfatizzata nel termine ombrello dell'Open science, dove trova, se possibile, ancora maggiore dignità accanto alla vasta gamma di componenti (open access publishing, open notebook, open peer review, citizen science, etc.) che supportano e caratterizzano una visione della scienza basata sui principi di trasparenza, collaborazione e integrità della ricerca. È evidente qui il collegamento con i quattro principi di Merton (1957), in particolare con quelli di comunalismo e scetticismo organizzato. Semplificando, il primo costituisce la premessa da cui scaturiscono gli altri principi mertoniani ed è strettamente collegato con le istanze di accesso aperto di tutti i prodotti della ricerca. Il secondo, applicato al contesto della validazione dei dati, è connesso alla peer review, quale strumento che certifica la validità del metodo e dei risultati.

Partendo dall'assunto che i dati rappresentano un primo significativo tassello nella costruzione del processo di ricerca e che la loro condivisione è il presupposto per il progresso della conoscenza, è lecito porsi ancora una volta la domanda di cosa significa rendere i dati accessibili alla luce dei cambiamenti tecnologici nel sistema della comunicazione scientifica. Per Lawrence (2011) la Pubblicazione dei dati (con la P maiuscola) costituisce la pratica di "rendere i dati il più possibile permanentemente disponibili su Internet" e nel contempo sottoporli ad un processo di creazione di metadati e di peer review (Mayernik et al. 2015) al fine di garantire qualità, trasparenza e riproducibilità.

Attualmente le principali fonti di pubblicazione[1] dei dati sono i data repository[2] e i data journal[3]. Entrambi diffondono i dati della ricerca allo scopo precipuo di renderli riutilizzabili e applicano forme specifiche di valutazione che mirano a verificare la loro affidabilità (*trustworthness*). Quando si instaura uno stretto legame tra repository e data journal (tramite doi bidirezionali e/o link persistenti), si crea un circolo virtuoso (Callaghan et al. 2014). Tale legame infatti rappresenta un importante incentivo per i produttori a rendere disponibili i propri dati: il data journal fornisce una chiara attribuzione del lavoro dando credito all'autore/i e/o istituzione, sottopone l'articolo al processo di peer review applicando determinati politiche editoriali e criteri di affidabilità. Inoltre e forse tra le più importanti conseguenze, il data journal rientra a pieno titolo nei tradizionali canali della comunicazione scientifica e, al pari degli altri tipi di pubblicazione, è accessibile, riutilizzabile e quindi potenzialmente citabile. In tal modo si assicura la priorità e il riconoscimento del

[1]In questo articolo il termine pubblicazione dei dati sottintende sempre un processo di validazione così come definito da Lawrence

[2]In questo articolo il termine "data repository" è usato nella sua accezione più ampia e comprende: i repository istituzionali, i repository disciplinari, i data centre e le infrastrutture di ricerca a livello internazionale.

[3]L'articolo si concentra in particolare sui data journal cosiddetti *puri*, in quanto pubblicano solo data paper (Candela et al. 2015). I data paper descrivono un dataset, fornendo dettagli sulla sua raccolta, elaborazione, calibrazione, software, formati di file, ecc. senza la necessità di nuove analisi o conclusioni innovative (Callaghan et al. 2013). Non vengono presi in considerazione tutte le altre tipologie di articoli o di documentazione supplementare all'articolo che pure sono stati precursori dei data journal.

produttore dei dati rafforzandone la reputazione. I data repository e i data journal, che attivano in modo coordinato un processo di scrutinio della qualità del dato, sostanziano dunque il principio di integrità della ricerca in quanto la rendono riproducibile e/o replicabile. Sono questi tutti elementi funzionali alla visione del sistema di scienza di Merton.

Ma cosa si intende per qualità del dato? Naturalmente entriamo in un terreno minato, ma cerchiamo di semplificare. Il RIN report (2008) associa la qualità al concetto di "fit for purpose" vale a dire la corrispondenza tra il dato e lo scopo per il quale è stato raccolto, quindi la congruenza tra i dati e i risultati ottenuti (sia positivi che negativi) che sostanziano il metodo della ricerca. A questo principio di carattere metodologico si affianca il requisito di fornire una descrizione appropriata del dato in modo da permetterne il controllo e la validazione da parte di altri ricercatori da un punto di vista del contenuto. L'altro requisito è collegato a tutti quegli elementi che dovrebbero rendere il dato idealmente recuperabile, accessibile e riutilizzabile e coinvolge, secondo gli autori del RIN, gli aspetti tecnici di creazione e cura del dato. Wang e Strong (1996) specificano che la qualità del dato è determinata dal "fitness for use by data consumers", attribuendo quindi un ruolo determinante a chi li utilizza. Se ne deduce che la validazione dei dati della ricerca costituisce un processo complesso, che viene reiterato in diverse fasi del processo di pubblicazione e che coinvolge professionalità e competenze specifiche, non ultime quelle dei potenziali utilizzatori, anch'essi coinvolti in un dialogo aperto che contribuisce all'avanzamento delle conoscenze.

Le riflessioni riportate in questo articolo fanno parte delle attività di analisi preliminari allo sviluppo di uno studio pilota previsto nel progetto europeo OpenUp (Opening up new methods, indicators and tools for peer review, dissemination of research results, and impact measurement[4]). Pertanto l'articolo, dopo aver presentato alcuni tra i principali aspetti del dibattito sulla validazione dei dati, fornisce una breve descrizione del progetto soffermandosi sui primi risultati e sulle considerazioni che faranno da guida per la scelta della comunità scientifica su cui testare l'applicabilità della peer review ai dati della ricerca in discipline afferenti alle scienze sociali. Il legame tra la validazione dei dati della ricerca e l'ethos mertoniano viene affrontato nelle considerazioni finali.

2. PEER REVIEW DEI DATI: UN DIBATTITO APERTO

La valutazione dei dati corrisponde, in gran parte, al processo di validazione reiterato durante l'intero ciclo di vita della ricerca. Parte già in fase di proposta, con la definizione di un *data management plan* che guida il ricercatore a raccogliere e gestire i dati secondo criteri di qualità. Viene applicata su un set coerente di risultati, ottenuti dall'elaborazione dei dati raccolti, al momento della sottomissione secondo criteri e procedure stabiliti da chi si assume il compito di pubblicare i risultati della ricerca. Infine viene attuata dai membri della comunità scientifica che accedono al dataset e lo riusano per produrre altri risultati. Il punto cruciale è quindi rappresentato dalla pubblicazione del dato che implica una valutazione della sua qualità, a differenza di quanto avviene nel caso della condivisione dei dati della ricerca, ad esempio in pagine web non sottoposte ad un processo di revisione formale. Tale valutazione costituisce un valore aggiunto per il potenziale utilizzatore (ne certifica l'affidabilità, *trust*) e nello stesso tempo dà credito accademico all'autore/i e/o istituzione che li hanno prodotti. Per questo la pubblicazione viene messa in relazione non solo con la peer review (tradizionale e/o aperta), ma anche con la citabilità del dataset, in quanto anch'essa costituisce una attribuzione di merito (positiva o negativa) da parte dei pari. Si delineano quindi due momenti essenziali di valutazione, quella effettuata in fase di pre-pubblicazione da esperti nella cura e gestione del dato e del dominio, e quella attuata in fase di post-pubblicazione dalla comunità di riferimento attraverso le citazioni (ma nel caso del dato ciò implica l'adozione di uno standard ancora da definire), ma anche utilizzando gli strumenti dell'altmetrics[5].

[4]OpenUp website http://openup-h2020.eu/
[5]https://www.altmetric.com/

In genere la maggior parte degli studiosi è concorde nell'affermare che il processo di peer review dei dati è più complesso e articolato rispetto a quello degli articoli scientifici. Esso richiede un consenso non ancora raggiunto, se non a livello di specifiche comunità, di quali aspetti del dato devono essere oggetto della validazione (Kratz and Strasser 2015). A ciò si aggiunge la complessità legata alle diverse tipologie di dati, creati e raccolti, nelle diverse discipline, per scopi differenti, con diversi metodi, strumenti e tecniche. Inoltre il dato è di per sé un oggetto dinamico continuamente aggiornabile e modificabile nel corso del processo di ricerca. Entrano in gioco quindi aspetti legati alla identificazione del dataset come prodotto della ricerca stabile, completo e permanente (Callaghan et al. 2012; Mayernik et al. 2015) che dovrebbe essere facilmente recuperabile per essere poi riutilizzato e quindi riproducibile. La possibilità di riprodurre e/o replicare i risultati di una ricerca, a partire dai dati raccolti, rappresenta infatti il fondamento del metodo scientifico su cui si basa il progresso della scienza. Ciò permetterebbe di *salire sulle spalle,* se non di giganti, di un corpus di risultati sui cui poi costruire nuove conoscenze (Merton 1991).

Come già accennato, si sta delineando un modello ideale di pubblicazione rappresentato dalla stretta collaborazione tra *trusted* data repository e data journal. I repository certificati (*trusted*) sono quelli che sviluppano chiare politiche di gestione dei risultati della ricerca, assicurano cura e conservazione a lungo termine dei dati, permettono l'accesso al dataset, implementano procedure di controllo di qualità, sviluppano strumenti di ricerca per recuperarli e forniscono statistiche d'uso (Callaghan et al. 2014; Whyte and Ball 2013). I data journal, a loro volta, in accordo con le proprie politiche editoriali, forniscono due tipi di linee guida: una per il produttore/autore dei dati e l'altra per i revisori. Le prime dovrebbero guidare il produttore/autore nella descrizione del dataset in termini di metodologia e protocollo utilizzato per raccogliere ed elaborare i dati, le seconde dovrebbero indicare ai revisori i criteri rispetto ai quali valutare il data paper.
In tale contesto il data journal, visto come tipologia di pubblicazione che promuove il riuso e la citazione, fa da ponte tra il dataset accessibile in un data repository certificato e l'articolo scientifico *tradizionale*, in quanto fornisce una documentazione dettagliata della metodologia e del processo di raccolta dei dati rimandando al set originario (tramite doi bidirezionali e/o link persistenti). Può inoltre far riferimento anche all'articolo *tradizionale* che invece si focalizza sull'interpretazione dei risultati. Il legame bidirezionale tra repository e data journal facilita l'identificazione e il recupero del dataset vero e proprio che è reso accessibile dal repository anche in termini di licenze e formati. Si noti inoltre che i requisiti di recuperabilità (nell'accezione di *findable* dei FAIR principles[6]) e accessibilità rappresentano le precondizioni necessarie per rendere il dato riutilizzabile, per questo gli elementi che istanziano tali requisiti nel dato e relativo metadato, sono anch'essi oggetto di controllo di qualità e anzi costituiscono il denominatore comune della maggior parte dei repository e data journal (Mayernik et al. 2015).

Numerosi e diversi sono gli altri aspetti del dato e del metadato che necessitano un controllo di qualità, in quanto entrano in gioco le specificità disciplinari, le metodologie e le tecniche di acquisizione. Molte le indicazioni teoriche (Costello and Wieczorekb 2014; Lawrence et al. 2011; Parsons and Fox 2013), ma altrettanto numerosi gli studi che analizzano le pratiche in uso e che indicano una diversità di approcci sia nelle pratiche di controllo di qualità dei repository (Assante et al. 2016) che nelle linee guida che dovrebbero guidare il revisore nel proprio lavoro (Candela et al. 2015; Carpenter 2017). Probabilmente, ciò dipende dal fatto che alcuni criteri di qualità del dato (completezza, autenticità, integrità, interoperabilità, riusabilità, etc.) possono essere esaminati da punti di vista diversi che dipendono in gran parte dal contesto in cui vengono validati i dati (Reilly et al. 2011) e dalle finalità specifiche di ciascun stakeholder coinvolto nel processo di pubblicazione. Valga a titolo esemplificativo il requisito del riuso, che viene comunemente identificato come una delle qualità fondamentali del dato.

[6]I FAIR principles sviluppati da FORCE11 sono quattro: Findable, Accessible, Interoperable, and Re-usable. https://www.force11.org/group/fairgroup/fairprinciples

In genere il riuso è visto sotto due accezioni che riguardano due aspetti tra loro correlati. Uno si riferisce più specificamente alla qualità del metadato che facilita il riuso in quanto permette di ricostruire il processo di raccolta ed elaborazione (ad esempio, corretta e dettagliata metadatazione, esplicita indicazione sulle licenze, uso di software non proprietario, etc.), l'altro è collegato alla valutazione del suo possibile impatto per la comunità di riferimento (FAIR principles). Entrambi gli aspetti sono oggetto di valutazione, se si vuole, semplificando, il primo riguarda gli aspetti tecnici del dato, il secondo quelli più propriamente scientifici (Callaghan et al. 2012). Tuttavia il riuso legato all'impatto rimane il più problematico da valutare a priori (non solo per i dati). In genere può essere confermato dal fatto stesso di essere citato una volta che il dato viene pubblicato e, così come per gli altri tipi di pubblicazione, indicazioni sulla sua rilevanza possono essere misurate calcolando il numero di citazioni ricevute. Ciò rientra tra le principali metriche di valutazione in fase di post-pubblicazione e, anche se anch'essa non è esente da critiche (H. F. Moed 2005, 2007; Bornmann and Daniel 2008) come del resto la peer review tradizionale (Lee et al. 2013; Nicholas et al. 2015), rimane pur sempre una delle più radicate forme di *dialogo* tra studiosi volte a dare credito e riconoscimento a chi ha prodotto il risultato della ricerca. Per questo motivo tener traccia di tale dialogo utilizzando metriche alternative, che forniscono importanti indicazioni sul numero di viste e download, commenti aperti in blog, etc., può contribuire non solo a confermare l'affidabilità del dato, ma anche fornire spunti e suggerimenti utili per l'avanzamento della ricerca.

3. IL PROGETTO OPENUP

OpenUP è un progetto finanziato dal programma europeo Horizon2020, nel quadro della tematica "Approcci innovativi nella disseminazione dei risultati e nella misurazione dell'impatto della ricerca scientifica". Il progetto ha preso avvio a giugno 2016 e include nove organizzazioni europee, tra università e centri di ricerca, di altrettanti paesi e coinvolge gruppi multidisciplinari di ricercatori, tra cui esperti di biblioteconomia, di tecnologie dell'informazione e della comunicazione, editori e scienziati sociali. OpenUP ricade tra le azioni di "Coordinamento e supporto alla ricerca" del programma quadro europeo finalizzate alla creazione di network, alla proposta di politiche e linee guida nonché allo sviluppo di piattaforme tecnologiche.

OpenUP prende in considerazione tre temi centrali per lo sviluppo della scienza aperta: l'open peer-review, la disseminazione innovativa dei risultati della ricerca e gli strumenti alternativi che misurano l'impatto della ricerca. Nella metodologia del progetto queste tematiche rappresentano i pilastri sui cui analizzare le trasformazioni nel sistema della comunicazione scientifica allo scopo di: 1) identificare meccanismi, processi e strumenti innovativi per la peer review applicata a tutti i risultati della ricerca (pubblicazioni, software e dati), 2) esplorare i meccanismi della disseminazione innovativa efficaci per le imprese, l'industria, il settore educativo e la società nel suo insieme e 3) analizzare un insieme di nuovi indicatori per la valutazione dell'impatto dei risultati della ricerca collegandoli ai canali per la disseminazione.

OpenUp utilizza una metodologia centrata sull'utente. Questo approccio metodologico non solo coinvolge tutti gli stakeholder (ricercatori, case editrici, enti che finanziano la ricerca, istituzioni, industria e il pubblico in generale) in una serie di workshops, conferenze e corsi di formazione, ma vuole testare i risultati acquisiti attraverso la realizzazione di sette studi pilota. Questi ultimi sono collegati ai tre pilastri e sono applicati ad alcune comunità e settori disciplinari: scienze umane e sociali, energia e scienze della vita. Per tutti gli studi pilota è stata definita una struttura simile e un quadro metodologico comune a livello di progetto. Tuttavia, date le differenze disciplinari e le pratiche specifiche delle varie comunità, ogni studio pilota ha elaborato propri criteri per analizzare in dettaglio le diverse realtà.
Nel primo anno del progetto, è stato ricostruito, per ogni tematica, lo stato dell'arte sulle diverse modalità di diffondere i risultati della ricerca, focalizzando l'attenzione su pratiche innovative (quali ad esempio il coinvolgimento di non esperti nei progetti di citizen science), nuovi strumenti

tecnologici (quali le piattaforme di open peer review o gli aggregatori che rilevano l'uso dei risultati della ricerca) e i nuovi canali di supporto alla comunicazione scientifica (ad esempio social media e strumenti di ricerca collaborativa). Inoltre è stata condotta un'indagine sugli atteggiamenti e percezioni dei ricercatori su tematiche correlate ai tre pilastri (Stančiauskas and Banelytė 2017). Focalizzandoci sui risultati inerenti la open peer review, si evidenzia che in generale i ricercatori si dichiarano abbastanza soddisfatti dell'attuale sistema di peer review, anche se emergono differenze tra discipline e livello di carriera (identificati in 4 tipologie: "leading researcher", established researcher, "recognised researcher" e "early stage researchers"). Ricercatori afferenti alle ICT e alle scienze sociali hanno percentuali più alte di insoddisfazione rispetto ad altre discipline e in genere sono i ricercatori più giovani che manifestano una visione critica rispetto alla peer review tradizionale. Infatti, i primi due livelli di carriera hanno espresso una maggiore propensione verso il rinnovamento rappresentato dagli strumenti della open peer review, giudicando gli strumenti tradizionali come non del tutto trasparenti o premiali per il merito. In genere, le principali critiche al sistema attuale di valutazione riguardano in particolare la qualità dei commenti dei revisori, la durata e la mancanza di trasparenza del processo. Per quanto riguarda la open peer review, i ricercatori esprimono forte perplessità nel rendere nota l'identità di revisori e autori così come nel rendere pubblici i commenti dei revisori. Si dichiarano invece favorevoli all'uso di piattaforme aperte di pubblicazione e alla partecipazione al processo di revisione di una platea più ampia di studiosi. La maggioranza è inoltre favorevole alla peer review dei dati.

Attualmente OpenUP sta avviando i sette studi pilota, collegati, come già detto, ai tre pilastri che hanno lo scopo di testare in specifici contesti i risultati ottenuti nella prima fase del progetto e di identificare strumenti e pratiche di supporto per la realizzazione degli obiettivi di una scienza aperta.

Lo studio pilota condotto dall'IRPPS intende indagare l'applicabilità dell'(open) peer review ai dati della ricerca nelle discipline afferenti alle scienze sociali. In particolare, lo scopo dello studio pilota è, da un lato, identificare i punti di forza e di debolezza nel processo di revisione e validazione dei dataset e, dall'altra, evidenziare le pratiche che facilitano la trasparenza del processo, la diffusione dei dati, la loro affidabilità e riuso. Per raggiungere tale obiettivo è necessario adottare una metodologia su cui basare i criteri di scelta della comunità oggetto dello studio, individuare le tipologie di attori e le relative modalità di analisi. La ricostruzione del contesto costituisce il primo passo per poter individuare una comunità rappresentativa nelle scienze sociali considerando che questo settore comprende diverse sotto-discipline, ognuna con proprie pratiche e tecniche di ricerca. Per questo motivo è necessario individuare le pratiche in uso per comunicare i risultati della ricerca, la propensione a condividere i dati e/o a riutilizzare quelli prodotti da altri, le infrastrutture di ricerca disponibili. Tale analisi, di cui si riporta una breve sintesi nel paragrafo successivo, è stata condotta attraverso l'analisi della letteratura ed ha inoltre considerato i principali sistemi di diffusione dei dati a livello internazionale. Alcuni criteri di selezione della comunità e modalità di sviluppo dello studio pilota sono naturalmente precondizione per verificare l'applicabilità della peer review ai dati della ricerca. Il primo è sicuramente legato alla necessità di individuare una comunità che si è fatta promotrice della condivisione e lo fa mettendo a disposizione i dati in modo gratuito. L'altro aspetto, anche motivato dall'analisi fin qui condotta, è legato alla necessità di analizzare sia la comunità che rende disponibili i dati sia quella che li usa. In tal modo si potranno esaminare modelli di condivisione e valutazione dei dati della ricerca in una prospettiva che mira sempre più a rendere il processo di pubblicazione come una pratica aperta di collaborazione e partecipazione allargata tra pari.

4. MODALITÀ DI PUBBLICAZIONE DEI DATI NELLE SCIENZE SOCIALI

In generale è difficile definire l'area delle scienze sociali poiché essa non viene considerata come un tutto unico ma include diverse famiglie di discipline con altrettante tradizioni storiche e culturali che a loro volta influenzano metodologie e tecniche di ricerca. La varietà dei dati che scaturiscono dai diversi approcci metodologici (qualitativi, quantitativi e quali-quantitativi) e tecniche di indagine (questionari, interviste, focus group, etc.) condizionano le modalità di descrizione del dato e del metadato così come le pratiche di diffusione e riuso (Curty 2016). Il Digital Curation Centre (DCC)[7] ha sviluppato lo standard Data Documentation Initiative (DDI)[8] per la descrizione dei dati della ricerca nelle scienze sociali, individuando uno schema di metadati e un relativo dizionario che facilitano l'accesso, la citabilità e il riuso. Inoltre il DDI è interoperabile con altri schemi di metadati bibliografici quali il Dublin Core[9], il Marc[10] e lo standard di scambio di dati statistici Statistical Data and Metadata eXchange (SDMX)[11] (Vardigan, Heus, and Thomas 2008). Tuttavia la diffusione del DDI è tuttora limitata a un numero ristretto di data repository.

In questo settore una porzione consistente di dati è rappresentata dai dati ufficiali o big data governativi (The Royal Society 2012) prodotti per scopi diversi da quelli strettamente scientifici (Borgman 2007). Gli Enti governativi e gli Istituti statistici nazionali anche in virtù delle recenti normative (si pensi in Italia alla legge sulla trasparenza, al Freedom act negli USA, al Freedom of information act in Gran Bretagna) hanno l'obbligo di rendere pubblici i dati da loro prodotti. Tali enti seguono procedure standardizzate per la raccolta e elaborazione dei dati (anche in materia di privacy) e forniscono in genere una documentazione dettagliata che descrive sia aspetti metodologici che tecnici del dataset messo a disposizione. Si tratta per lo più di indagini ripetute a intervalli regolari (tipicamente i censimenti), o su temi specifici che tendono a rilevare atteggiamenti e comportamenti della popolazione rispetto a determinati fenomeni (si pensi alle indagini multiscopo dell'Istat o a quelle europee dell'Eurobarometer). I dati prodotti dagli Istituti di statistica nazionali e internazionali sono tra le principali fonti utilizzate dagli scienziati sociali per produrre nuove ricerche. Fonti centralizzate di dati, quali ad esempio il UK Data Archive[12], forniscono accesso sia a dati ufficiali che a quelli prodotti per scopi scientifici ed applicano tecniche simili di validazione che supportano l'affidabilità del dato.

In genere, si può affermare che il modello di pubblicazione del dato nelle scienze sociali rimane prevalentemente ancorato a quello di diffusione attraverso i data repository e/o le tradizionali riviste scientifiche, alcune delle quali richiedono documentazione supplementare per la descrizione dei dati. Attualmente esiste un solo data journal nel settore delle scienze sociali: "Research Data Journal (RDJ)"[13], creato dal Data Archiving and Network Services (DANS)[14] nel 2016 con lo scopo di aumentare la visibilità dei dati depositati nell'archivio e di fornirne una documentazione più ampia e dettagliata. La rivista segue il modello dei data journal in quanto assegna il doi all'articolo e lo collega a quello del dataset depositato in uno degli archivi consigliati. Il suo sviluppo recente non permette attualmente un'analisi comparativa con esperienze di altri settori disciplinari, ma costituisce senz'altro un prodotto editoriale interessante da monitorare in futuro.

I repository nelle scienze sociali rappresentano quindi le principali fonti di accesso ai dati. La qualità del dato è pertanto direttamente proporzionale alla qualità del repository. Più il repository ha politiche trasparenti e adotta procedure rigorose nel processo di cura e archiviazione del dato, più i dati pubblicati saranno di alta qualità scientifica e quindi affidabili e riutilizzabili. Tale relazione è

[7]Digital Curation Centre http://www.dcc.ac.uk/

[8]Data Documentation Initiative http://www.dcc.ac.uk/resources/metadata-standards/ddi-data-documentation-initiative

[9]Dublin Core http://dublincore.org/

[10]Marc standards https://www.loc.gov/marc/

[11]Statistical Data and Metadata eXchange https://sdmx.org/

[12]UK Data Archive http://www.data-archive.ac.uk/

[13]Research Data Journal http://dansdatajournal.nl/ddj/

[14]Data Archiving and Network Services https://dans.knaw.nl/en

ben rappresentata dalla piramide dei dati del rapporto della Royal Society (2012) in cui il valore del dato cresce in relazione alle pratiche con cui i dati vengono gestiti e conservati. La piramide presenta quattro livelli di tipologie di data repository, alla base si trovano le collezioni individuali messe a disposizione volontariamente dai produttori dei dati ad esempio nella propria pagina web, al livello successivo ci sono repository istituzionali che gestiscono i dati delle proprie ricerche. In questo caso, considerato che spesso i repository istituzionali gestiscono più tipologie di pubblicazione, le procedure di validazione risultato spesso eterogenee. Ai livelli più alti della piramide ci sono i data center in genere centralizzati in ambito nazionale, mentre l'apice è costituito dalle infrastrutture di ricerca a livello internazionale. Soffermandoci solo sui due livelli superiori, nelle scienze sociali ci sono importanti data center nazionali, quali il Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS)[15] e Swedish National Service (SNS) [16] i già menzionati UK Data Archive, il DANS, mentre tra i consorzi internazionali vanno ricordati a livello europeo il Consortium of European Social Science Data Archives (CESSDA Eric)[17] e a livello internazionale l'Interuniversity Consortium for Political and Social Research (ICPSR)[18]. I consorzi oltre a fornire un punto di accesso integrato ai dati, sviluppano e coordinano le iniziative su standard, protocolli e best practice a sostegno delle attività di gestione e distribuzione dei dati. Si tratta di *trusted* repository che applicano controlli di qualità attraverso politiche e procedure trasparenti, in alcuni casi basate sulle linee guida elaborate dal Data Seal of Approval[19], una certificazione internazionale di qualità dei repository. Tale certificazione può essere utilizzata dagli editori commerciali nella selezione dei repository consigliati quando si richiede all'autore di depositare il dataset.

Per garantire ad esempio la recuperabilità e la riusabilità del dato alcuni repository come ad esempio l'ICPSR, il DANS e il UK Data Archive utilizzano lo standard di metadati DDI per descrivere le collezioni dei dati. La descrizione del dataset attraverso questo schema di metadati permette di ricostruire il contesto del dato. Infatti oltre alle informazioni sul dataset (ad es. tipo di dati, variabili etc.) vengono inserite quelle che descrivono lo scopo dello studio, le metodologie e le procedure utilizzate fornendo in tal modo gli elementi indispensabili per il riuso del dato (Van den Eynden and Corti 2017). Per rendere i dati accessibili**,** tenendo presente la legislazione in materia sia di privacy sia di proprietà intellettuale, vengono fornite in genere delle linee guida su come anonimizzare i dati e viene suggerito l'uso di licenze di Creative Commons[20]. Ad esempio il DANS propone per i propri dati lo slogan "accessible whenever possible, protected if necessary"[21]. Inoltre sono state stabilite precise politiche a favore della citabilità del dato attraverso prima di tutto l'attribuzione del doi. Alcuni repository proprio per la natura dinamica dei dati e/o per rendere evidente l'intero ciclo di vita del dato attribuiscono il doi non al singolo file ma alla collezione, forniscono gli elementi obbligatori per una corretta citazione e a volte ne consigliano uno specifico formato (ad esempio l'APA per UK Data Archive). Alcuni repository hanno implementato strumenti che ritracciano l'uso dei dati in termini di conteggio delle viste e download come ad esempio le statistiche fornite dall'ICPSR, dal UK Data Archive e dal DANS. Inoltre il DANS ha realizzato uno studio pilota per valutare il "fit for reuse" dei propri dati chiedendo direttamente agli utilizzatori di valutarne la qualità (Grootveld and Van Egmond 2012).

[15]GESIS - Leibniz-Institut für Sozialwissenschaften in Mannheim https://www.gesis.org/en/home/

[16]Swedish National Service https://snd.gu.se/en

[17]Consortium of European Social Science Data Archives https://www.cessda.eu/

[18]Interuniversity Consortium for Political and Social Research https://www.icpsr.umich.edu/icpsrweb/

[19]Data Seal of Approval https://www.datasealofapproval.org/en/

[20]Licenze Creative Commons http://www.creativecommons.it/

[21]https://dans.knaw.nl/en/about/organisation-and-policy/legal-information/property-rights-statement

Infine è interessante menzionare il progetto "Data impact blog"[22] della UK Data Service[23] che attraverso le discussioni nel blog vuole incoraggiare il dibattito, condividere esperienze e best practice, mantenere la comunità aggiornata sulle problematiche emergenti.

5. CONSIDERAZIONI FINALI

Carpenter (2017) afferma che il numero dei dataset disponibili è aumentato del 400% dal 2011 al 2015 e alcune indagini (Reilly et al. 2011; Jeng et al. 2016) sembrano confermare un incremento dei dati sottomessi ai repository. Ciò avviene soprattutto quando sono gli editor a richiedere l'accesso al dato prima della valutazione tra pari. Non tutti concordano su un incremento così rilevante dei dati disponibili (Alsheikh-Ali et al. 2011), tuttavia è evidente che le diffidenze dei ricercatori rispetto alla condivisione dei dati, (Kratz and Strasser 2015; Swan and Brown 2008; Tenopir et al. 2011), vengono via via superate quando le spinte al cambiamento sono sostenute da una pluralità di attori (si pensi alle politiche dell'European Commission High Level Expert Group on Scientific Data 2010; del National Institute of Health 2016; della National Science Foundation 2011; dell'OECD 2007). Ciò dimostra, come già affermato da Merton, che è necessario che le norme individuali o di comunità trovino una corrispondenza con quelle dell'istituzione, intesa in senso lato. Solo questa alleanza permette di consolidare buone pratiche, nel nostro caso quelle della condivisione dei dati, che vengono *universalmente* riconosciuti come bene comune (Fecher et al. 2017; Merton 1973).

Il recente affermarsi dei data journal va in questa direzione, in quanto essi propongono un modello di stretta collaborazione tra gli editori dei data journal e i gestori dei data repository accomunati dall'intento di rendere i dati validati e accessibili alla comunità scientifica.

L'affermarsi di nuovi modelli di pubblicazione crea un effetto domino sulle diverse componenti del sistema scienza. In genere questi nuovi modelli sono proposti e sperimentati all'interno di specifiche comunità scientifiche che, in base alle loro pratiche di ricerca, si sono date regole comuni per la condivisione dei dati sviluppando nello stesso tempo strumenti quali piattaforme, standard di metadati, protocolli di validazione, etc. Si pensi al ruolo che hanno avuto i fisici nello sviluppo del modello di ArXiv come antesignano dell'open access o a quello che attualmente svolgono le comunità afferenti alle scienze della terra nel proporre il modello dei data journal[24], su cui vengono sperimentati anche diverse modalità di open peer review. Tali esperienze possono diventare un punto di riferimento per altre comunità e magari essere recepite come buone pratiche anche a livello istituzionale. Il graduale affermarsi del binomio data journal trusted repository ha permesso di analizzare sotto una diversa prospettiva i criteri di valutazione sia pre che post-pubblicazione, le modalità di accesso e non ultimi la formulazione di standard per la citazione dei dati.

Questo articolo ha analizzato le problematiche di validazione del dato privilegiando gli aspetti che lo rendono riproducibile e replicabile, aspetti chiave in una visione di integrità della scienza basata sul "dubbio sistematico" che si riflettono nei principi mertoniani. Tale integrità trova nella citazione una ulteriore forma di validazione che, come abbiamo spesso sottolineato, sottintende il riconoscimento dell'autore e l'attribuzione di priorità e pertanto, come afferma Merton, rappresenta l'unica moneta della scienza.

BIBLIOGRAFIA

[22]Data impact blog http://blog.ukdataservice.ac.uk/

[23]UK Data Service https://www.ukdataservice.ac.uk/

[24]Earth System Science Data https://www.earth-system-science-data.net/; Geoscience Data Journal http://rmets.onlinelibrary.wiley.com/hub/journal/10.1002/(ISSN)2049-6060/about/author-guidelines.html

Alsheikh-Ali Alawi A., Waqas Qureshi, Mouaz H. Al-Mallah, and John P. A. Ioannidis. 2011. "Public Availability of Published Research Data in High-Impact Journals." Edited by Isabelle Boutron. *PLoS ONE* 6 (9). Public Library of Science: e24357. doi:10.1371/journal.pone.0024357.

Assante Massimiliano, Leonardo Candela, Donatella Castelli, and Alice Tani. 2016. "Are Scientific Data Repositories Coping with Research Data Publishing?" *Data Science Journal* 15: 1–24. doi:10.5334/dsj-2016-006.

Borgman Christine L. 2007. *Scholarship in the Digital Age : Information, Infrastructure, and the Internet*. MIT Press.

Bornmann Lutz, and Hans Dieter Daniel. 2008. "What Do Citation Counts Measure? A Review of Studies on Citing Behavior." *Journal of Documentation* 64 (1). Emerald Group Publishing Limited: 45–80. doi:10.1108/00220410810844150.

Callaghan Sarah, Steve Donegan, Sam Pepler, Mark Thorley, Nathan Cunningham, Peter Kirsch, Linda Ault, et al. 2012. "Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres." *International Journal of Digital Curation* 7 (1): 107–13. doi:10.2218/ijdc.v7i1.218.

Callaghan Sarah, Fiona Murphy, Jonathan Tedds, Rob Allan, John Kunze, Rebecca Lawrence, Matthew S. Mayernik, and Angus Whyte. 2013. "Processes and Procedures for Data Publication: A Case Study in the Geosciences." *International Journal of Digital Curation* 8 (1): 193–203. doi:10.2218/ijdc.v8i1.253.

Callaghan Sarah, Jonathan Tedds, John Kunze, Varsha Khodiyar, Rebecca Lawrence, Matthew Mayernik, Fiona Murphy, Timothy Roberts, and Angus Whyte. 2014. "Guidelines on Recommending Data Repositories as Partners in Publishing Research Data." *International Journal of Digital Curation* 9 (1): 152–63. doi:10.2218/ijdc.v9i1.309.

Callaghan Sarah, Jonathan Tedds, Rebecca Lawrence, Fiona Murphy, Timothy Roberts, and Will Wilcox. 2014. "Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples." *International Journal of Digital Curation* 9 (1): 164–75. doi:10.2218/ijdc.v9i1.310.

Candela Leonardo, Donatella Castelli, Paolo Manghi, and Alice Tani. 2015. "Data Journals: A Survey." *Journal of the Association for Information Science and Technology* 66 (9): 1747–62. doi:10.1002/asi.23358.

Carpenter Todd A. 2017. "What Constitutes Peer Review of Data: A Survey of Published Peer Review Guidelines," April. http://arxiv.org/abs/1704.02236.

Costello Mark J., William K. Michener, Mark Gahegan, Zhi-Qiang Zhang, and Philip E. Bourne. 2013. "Biodiversity Data Should Be Published, Cited, and Peer Reviewed." *Trends in Ecology & Evolution* 28 (8): 454–61. doi:10.1016/j.tree.2013.05.002.

Costello Mark J., and Wieczorekb John. 2014. "Best Practice for Biodiversity Data Management and Publication." *Biological Conservation* 173 (May). Elsevier: 68–73. doi:10.1016/J.BIOCON.2013.10.018.

Curty Renata Gonçalves. 2016. "Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study." *International Journal of Digital Curation* 11 (1): 96–117. doi:10.2218/ijdc.v11i1.401.

Eynden Veerle Van den, and Louise Corti. 2017. "Advancing Research Data Publishing Practices for the Social Sciences: From Archive Activity to Empowering Researchers." *International Journal on Digital Libraries* 18 (2). Springer Berlin Heidelberg: 113–21. doi:10.1007/s00799-016-0177-3.

Fecher Benedikt, Sascha Friesike, Marcel Hebing, and Stephanie Linek. 2017. "A Reputation Economy: How Individual Reward Considerations Trump Systemic Arguments for Open Access to Data." *Palgrave Communications* 3 (June). Nature Publishing Group: 17051. doi:10.1057/palcomms.2017.51.

Grootveld Marjan, and Jeff Van Egmond. 2012. "Peer-Reviewed Open Research Data: Results of a

Pilot." *The International Journal of Digital Curation International Journal of Digital Curation* 7 (72): 81–91. doi:10.2218/ijdc.v7i2.231.
Jeng Wei, Daqing He, and Jung Sun Oh. 2016. "Toward a Conceptual Framework for Data Sharing Practices in Social Sciences: A Profile Approach. In the Proceedings of the ASIS&T 2016 Annual Meeting." Wiley. http://d-scholarship.pitt.edu/31883/.
Kratz John Ernest, and Carly Strasser. 2015. "Researcher Perspectives on Publication and Peer Review of Data." *PLoS ONE*. doi:10.1371/journal.pone.0117619.
Lawrence Bryan, Catherine Jones, Brian Matthews, Sam Pepler, and Sarah Callaghan. 2011. "Citation and Peer Review of Data: Moving Towards Formal Data Publication" 6 (2). http://dx.doi.org/10.2218/ijdc.v6i2.205
Lee Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in Peer Review." *Journal of the American Society for Information Science and Technology* 64 (1): 2–17. doi:10.1002/asi.22784.
Mayernik Matthew S., Sarah Callaghan, Roland Leigh, Jonathan Tedds, and Steven Worley. 2015. "Peer Review of Datasets: When, Why, and How." *Bulletin of the American Meteorological Society*. doi:10.1175/BAMS-D-13-00083.1.
Merton Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22 (6). American Sociological Association: 635. doi:10.2307/2089193.
Merton Robert K. 1973. *The Sociology of Science : Theoretical and Empirical Investigations*. University of Chicago Press.
Merton Robert K. 1991. *Sulle spalle dei giganti*. Bologna : Il Mulino
Moed Henk F. 2005. *Citation Analysis in Research Evaluation*. Springer.
Moed Henk F. 2007. "The Effect of 'open Access' on Citation Impact: An Analysis of ArXiv's Condensed Matter Section." *Journal of the American Society for Information Science and Technology* 58 (13). Wiley Subscription Services, Inc., A Wiley Company: 2047–54. doi:10.1002/asi.20663.
Nicholas David, Anthony Watkinson, Hamid R. Jamali, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. 2015. "Peer Review: Still King in the Digital Age." *Learned Publishing* 28 (1): 15–21. doi:10.1087/20150104.
Parsons Mark A., and P A Fox. 2013. "Is Data Publication the Right Metaphor?" *Data Science Journal* 12 (0). Ubiquity Press: WDS32-WDS46. doi:10.2481/dsj.WDS-042.
Parsons Mark A., Ruth Duerr, and Jean Bernard Minster. 2010. "Data Citation and Peer Review." *Eos* 91 (34): 297–98. doi:10.1029/2010EO340001.
Reilly Susan, Wouter Schallier, Sabine Schrimpf, Eefke Smit, and Max Wilkinson. 2011. "Report on Integration of Data and Publications," January. doi:10.5281/ZENODO.8307.
Stančiauskas Vilius, and Viltė Banelytė. 2017. "Openup Survey On Researchers' Current Perceptions And Practices In Peer Review, Impact Measurement And Dissemination Of Research Results," January. doi:10.5281/ZENODO.556157.
Swan Alma, and Sheridan Brown. 2008. "To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs. A Report Commissioned by the Research Information Network." s.n. https://eprints.soton.ac.uk/266742/.
Tenopir Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." Edited by Cameron Neylon. *PLoS ONE* 6 (6). Public Library of Science: e21101. doi:10.1371/journal.pone.0021101.
The Royal Society. 2012. "Science as an Open Enterprise" London . https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/.
Vardigan, Mary, Pascal Heus, and Wendy Thomas. 2008. "Data Documentation Initiative: Toward a Standard for the Social Sciences." *International Journal of Digital Curation* 3 (1): 107–13. doi:10.2218/ijdc.v3i1.45.
Whyte Angus, and Alex Ball. 2013. "Data Publishing, Peer Review and Repository Accreditation:

Everyone a Winner? Report from the PREPARDE Project IDCC Workshop Annex to PREPARDE Deliverable D5.1 Report on Requirements for Data Centre Accreditation."